Signal Quality Detection Towards Practical Non-Touch Vital Sign Monitoring

Zongxing Xie
Stony Brook University
zongxing.xie@stonybrook.edu

Bing Zhou
IBM Research
bing.zhou@ibm.com

Fan Ye
Stony Brook University
fan.ye@stonybrook.edu

ABSTRACT

Non-touch vital sign sensing is gaining popularity because it does not require users' cooperative efforts (e.g., charging, wearing) thus convenient for longitudinal monitoring. In recent radio-based heart and respiration rate (HR and RR) sensing using Wi-Fi, millimeter wave (mmWave), or ultra-wideband (UWB), inevitable user movements or background moving objects cause large disturbances to the much weaker respiratory and heart signals. Such "corrupted" signals must be detected and excluded to avoid making erroneous measurements. Despite several attempts, reliable signal quality detection (SQD) remains unresolved. In this paper, we spent over 80 hours to manually examine 50268 data samples collected from 8 participants. We find that heart and respiration signals are not always simultaneously available, which breaks an important assumption in prior work. We propose a *2-bit SQD* to classify their "availability" separately. We further quantify the contributions of and correlation among a comprehensive set of features in both time and frequency domains, and use a forward selection strategy to identify an optimal and much smaller feature set for multiple common classification algorithms. Extensive experiments show that our 2-bit SQD achieves 91/95% precision, 88/91% recall in detecting available RR/HR signals, as compared to a flat spectrum detector (FSD) [3] and a spectrum-averaged harmonic path detector (SHAPA) [24] in prior work, and reduces the 80-percentile RR/HR errors from 10/18 bpm to 3.5/4.0 bpm, 3~4 fold reductions.

CCS CONCEPTS

• **Applied computing → Bioinformatics**; **Health informatics**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Signal quality detection, non-touch vital sign monitoring, longitudinal in-home data collection

ACM Reference Format:
Zongxing Xie, Bing Zhou, and Fan Ye. 2021. Signal Quality Detection Towards Practical Non-Touch Vital Sign Monitoring . In *12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '21), August 1–4, 2021, Gainesville, FL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3459930.3469526

1 INTRODUCTION

Basic vital signs such as heart and respiration rates (HR, RR) play an important role in human health status assessment [5]. Continuous vital signs data collected in individuals' home environment are invaluable for detecting potential health problems, and provide insights for managing medication plans, especially for older adults who face a myriad of chronic diseases and health conditions. However, continuous vital sign monitoring remains difficult to achieve.

Such longitudinal in-home monitoring requires robust and passive sensing. Wearables (e.g., Apple Watch, Fitbit) require frequent charging and wearing, thus difficult to maintain compliance, especially among physically and cognitively challenged older adults. Recent radio-based sensing designs [11] exemplified by Wi-Fi [19, 32], frequency-modulated continuous wave (FMCW) [3] and UWB [24] solutions require zero cooperative efforts from users, thus holding promise for longitudinal in-home monitoring.

Although such work has demonstrated feasibility in lab environments (e.g., the subject remains stationary and clear from moving objects), robustness in real daily environments faces great challenges from signal corruption. Inadvertent body motion, moving objects, multi-path reflections etc. can cause large disturbances, making the signal corrupted beyond recognition by even well-trained human experts. Such corrupted signals cause large errors and pollute longitudinal data, thus must be detected and excluded.

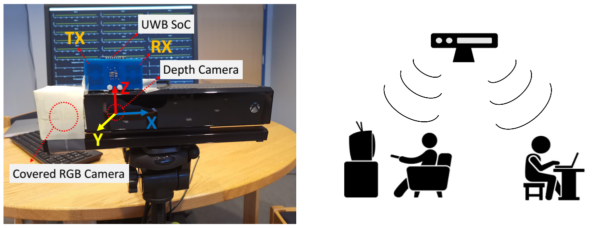
Existing SQDs [3, 11, 35] rely on simple heuristics and usually use a single feature such as spectral energy or temporal waveform for classification. Such naive methods may ignore available signals, or include corrupted ones in face of complex dynamics such as non-linear channels, heart-respiration intermodulations [18]. They also assume that the presence of RR and HR are always simultaneous, i.e., either they are both "available" or both "corrupted." However, by careful manual examination of large amounts of data samples, we find that about 29% of the time, one signal can be present while the other is corrupted. Thus a single binary detector may incorrectly throw out available signals, or include corrupted ones.

To tackle this problem, we propose a 2-bit SQD that examines the availability of HR, RR individually. We create a comprehensive set of signal features in both time and frequency domains, rigorously quantify their uniqueness and importance, and identify a small optimal feature set for each of several common classification algorithms. To evaluate the proposed 2-bit SQD, we conduct extensive

This work is supported in part by NSF grants 1652276, 1951880, 2028952. Bing Zhou contributed to this work when he was affiliated with Stony Brook University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

BCB '21, August 1–4, 2021, Gainesville, FL, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8450-6/21/08…$15.00
https://doi.org/10.1145/3459930.3469526



(a) System layout.

(b) In-home sedentary scenarios.

Figure 1: Our non-contact vital sign monitoring system consists of a commercial UWB SoC as the RF front end and a co-located Kinect XBox (with RGB camera covered for privacy) whose depth sensor provides human detection and location. Together they enable automated sensing without users' cooperative efforts.

experiments in home environments. Figure 1 shows the system and scenarios for non-contact vital sign monitoring adopted in this paper. We use a commercial UWB system on a chip (SoC) as the radio frequency (RF) front end, which emits impulse radio waves and received echoes from the field of view (FOV). Our system targets the scenarios where users are in sedentary activities, during which they stay quasi-stationary and their vital signs can be demodulated from the received radio echoes, while the disturbed signal during the non-stationary period will be detected and excluded by SQD. Our experimental results show that compared to existing single feature based detectors, our 2-bit SQD is more robust, and reduces end-to-end RR/HR errors at 80-percentile by 3∼4 folds.

Specifically, we make the following contributions in this work:

- We conduct over 80 hours' careful manual examination on 50268 samples in 840 minutes data from 8 participants. We find that 29% of the time the availability of HR, RR signals are not the same, breaking an implicit assumption made in all prior work and necessitating 2-bit SQD that treats HR, RR separately for reliability.
- We create a comprehensive sets of 29 signal features in both time and frequency domains, quantify their correlations to identify redundant features, and measure their individual contributions to detection. Using a forward selection strategy, we find for each of several common classifiers, a much smaller feature subset (usually 3–10 features) can achieve optimal detection.
- We evaluate the end-to-end vital sign estimation performance using our 2-bit SQD in typical home environments. We achieve 91/95% precision and 88/91% recall for RR/HR on previously unseen subjects' data, and 3.5/4.0 *bpm* 80-percentile error, 3∼4 fold reduction compared to over 10/18 bpm errors at 80-percentile using existing detectors.

2 RELATED WORK

Traditional measurement technologies of vital signs (heartbeat and respiration rates) are contact-based, including those using electrocardiogram (ECG) [26], photoplethysmogram (PPG) [30] and ballistocardiogram (BCG) [13]. Wearables usually leverage such technologies, and require constant wearing, charging efforts, difficult to comply over long time, especially for older adults. We use a non-contact vital signs sensing system with radio-based techniques targeting quasi-stationary settings. It requires no cooperative efforts (e.g., putting on a device), thus more suitable for longitudinal in-home monitoring. Our main focus in this work is SQD, how to reliably detect and exclude "corrupted" signals to improve robustness of vital signs measurements.

Morphological features (such as systolic peaks, maximum slopes) modulated by cardiovascular activities and extracted from ECG, PPG or BCG signals have demonstrated reliable heart rate estimation [13, 26, 30]. However, the radio perceived physiological signal is modulated by chest wall movements, a combination of respiratory and heart activities, and the waveform is dominated by the much stronger respiration signal. Thus such morphology-based detection is not applicable to radio-based vital signs sensing.

Different methods have been proposed to detect corrupted signals. Threshold-based methods reject the motion artifact according to motion level, quantified based on signal patterns in the time domain [19, 34] or frequency domain [3, 11]. A representative one is the flat spectrum detector (FSD [3]), which uses the peak-to-average ratio of the frequency spectrum to indicate the motion level. Signals disturbed by motion artifacts will have relatively small peak-to-average ratio as the energy is spread over the whole spectrum. If the motion level is greater than a predetermined threshold, the signal is excluded. However, we observe such fixed thresholds are unreliable to precisely reject motion artifacts. SHAPA [24] seeks to detect the availability of vital signs based on a heuristic that the presence of multiple orders of harmonics indicates vital signs are available. Although a strong evidence, SHAPA may exclude signals where only vital signs fundamental components are present and can be extracted, resulting in low recall. We also notice that all such work implicitly assumed that respiration and heartbeat signals are either both available or both unavailable, which we debunk by careful manual examination. We further devise separate SQDs for respiration and heartbeat, and identify optimal feature sets using rigorous feature selection to greatly improve the reliability of vital signs estimation.

3 BACKGROUND

In this section, we describe the background of signal modeling in UWB-based vital signs extraction and formulate a scoped problem.

3.1 UWB-based Vital Signs Extraction

The displacement of the chest wall, as a result of the combination of heartbeat and respiration, can be extracted from received UWB signals and modeled in the equation:

$$\begin{aligned} d(t) &= d_0 + D(t) \\ &= d_0 + d_r \sin(2\pi f_r t) + d_h \sin(2\pi f_h t), \end{aligned} \tag{1}$$

where d_0 is the nominal distance between the UWB sensor and the targeted chest wall; d_r and d_h are the chest displacement amplitudes, and f_r and f_h the rates of respiration and heartbeat, respectively.

Changing displacement of the chest wall causes the delay of the reflected UWB signal $\tau_D(t) = 2d(t)/c$ to change. Therefore, the phase of the reflected signal is modulated by such periodic displacement, and can be modeled as:

$$\phi(t) = \phi_0 + \phi_D(t), \tag{2}$$

where ϕ_0 is the initial phase of the signal reflected from a nominal distance d_0, $\phi_D(t) = 2\pi f_c D(t)/c$ is the phase modulated by the physiological movements, and f_c is the center frequency of the UWB signal. By extracting the frequency components from the modulated phase, we are able to extract RR/HR.

3.2 Problem Formulation

As derived in §3.1, the physiological signal can be obtained via phase demodulation. The signal is inherently periodic because both heartbeat and respiration are periodic activities. Therefore, vital signs can be extracted with existing spectral analysis techniques (e.g., Fourier Transform). While the approach appears straightforward, reliable vital sign estimation is greatly complicated by inadvertent movements and non-linear channels in the real world.

Since the human body is relatively large compared to the wavelength of UWB signals (a few centimeters), it is necessary to model human body as a sum of multiple scatterers rather than a single point scatterer. The perceived signal is in the superposition of echoes bounced off every body part. Small motions of any body parts in proximity of the chest wall could lead to dominating disturbance in the physiological signals, resulting in corrupted signals. SQD is a must to exclude erroneous vital sign estimations from such corrupted signals.

Real world signal channels are often non-linear, adding another challenge for robust SQD. The perceived phase under such non-linear channels can be more complex than $\phi_D(t) = 2\pi f_c D(t)/c$. A complex non-linear signal can be approximated by its Taylor series as follows:

$$\phi_D(t) = \frac{2\pi f_c}{c}(a_1 D(t) + a_2 D^2(t) + a_3 D^3(t) + \ldots), \tag{3}$$

where a_i is the coefficient of the i-th order term. The higher order terms result in intermodulation products between heartbeat and respiratory signals. Take the second order item for example:

$$\begin{aligned} D^2(t) =& d_r^2 \sin(2\pi f_r t)^2 + d_h^2 \sin(2\pi f_h t)^2 \\ &+ 2 d_r d_h \sin(2\pi f_r t)\sin(2\pi f_h t). \end{aligned} \tag{4}$$

The products of trigonometric functions are equivalent to the following sums, according to trigonometric formulas:

$$\begin{aligned} \sin(2\pi f_r t)^2 =& \frac{1}{2} - \frac{1}{2}\cos(\underline{4\pi f_r t}), \\ \sin(2\pi f_h t)^2 =& \frac{1}{2} - \frac{1}{2}\cos(\underline{4\pi f_h t}), \\ \sin(2\pi f_h t)\sin(2\pi f_r t) =& \frac{1}{2}\cos(\underline{2\pi(f_h - f_r)t}) \\ &- \frac{1}{2}\cos(\underline{2\pi(f_h + f_r)t}). \end{aligned} \tag{5}$$

As indicated in the underlined items in above equations, *intermodulation* manifests as new spectral components of linear combinations of heart and respiratory rates (i.e., $\{m f_h \pm n f_r | m, n \in \mathbb{N}_0\}$), at frequencies close to heart, respiration rates, making it difficult to tell which are true vital signs.

Simply including the corrupted signal for vital signs extraction will introduce erroneous measurements, thus subsequent meaningless and misleading analysis results. Therefore, robust SQD is a critical part for reliable vital sign monitoring. In this paper, we propose a robust 2-bit SQD to exclude corrupted signals thus erroneous vital sign measurements.

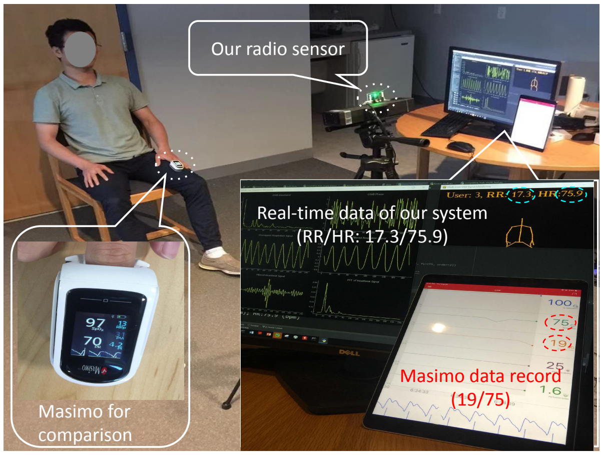


Figure 2: A typical setup for non-contact vital sign monitoring and data collection. The subject sits on the chair at a distance from the radio sensor. The fingertip Masimo pulse oximeter [1] is used to provide ground truth for comparison.

4 2-BIT SIGNAL QUALITY DETECTION

In this section, We first describe our data collection, then descrbie the design of the 2-bit SQD.

4.1 Data Collection

We use the following system and protocol for data collection to create the datasets for the design and evaluation of 2-bit SQD.

4.1.1 Wireless Physiological Sensing System. We use a COTS UWB sensor (XeThru X4M03 [2]) as the RF front end, which transmits impulse radio waves within the frequency band 7.25–10.2 GHz (centered at 8.75 GHz). The UWB sensor produces 10 frames per second (fps), and each frame includes samples of the echoes reflected from objects within a range of 10 meters. To detect the distance to the human body thus "zooming in" corresponding signals, we use Kinect XBox's depth sensor, which integrates a human body pose recognition model [28] to recognize and localize human bodies present in the field of view. Both sensors stream data to the same backend PC via serial ports for further processing. Alongside the radio sensing system, we use a FDA approved medical device, Masimo Pulse Oximeter [1], as the physiological reference system to obtain ground truth.

Figure 2 shows a typical setup for wireless physiological monitoring. The UWB sensor sits on top of a Kinect XBox One sensor for wireless physiological sensing, and the Masimo device is placed on the fingertip to measure instantaneous heartbeat and respiratory rates as the ground truth. We use a sliding window of 30 seconds at 1 second increments to generate data samples, and align each sample with the ground truth according to the timestamp.

4.1.2 Protocol. We conduct data collection in home environments. In total 11 participants contribute to data collection, following a pre-established protocol that protects the anonymity of the participants. During data collection, we ask the participants to freely perform three common sedentary activities (i.e., reading, typing and resting on the chair), at varying distances (1-3 meters) and orientations (0-90 degrees) relative to the sensing system.

We collect three data sets. The first data set (denoted as $D1$) has in total 50268 samples. It is collected with 8 out of 11 participants over two weeks for preliminary tests of SQD. Each participant spends about 1 hour for data collection at 9 different combinations of distances and orientations in sedentary activities. The second data set (denoted as $D2$) has in total 4877 samples. It is collected with the rest 3 participants to evaluate the robustness of the SQD model against unseen subjects. Each participant spends about 30 minutes for data collection at randomly selected distances and orientations in sedentary activities. The third data set (denoted as $D3$) has in total 26450 samples. It is collected with 2 same participants in D1. Each participant spends about 30 minutes per day intermittently over 2 months. This data set is to evaluate the end-to-end performance for longitudinal vital sign monitoring.

4.2 Signal Quality Detection

We first inspect the data characteristics, then devise a set of optimal features and train models for 2-bit SQD.

4.2.1 Data Characteristics. To understand the data characteristics, we first manually label the availability of each samples in $D1$. We have one expert to annotate the availability of each sample for respiration and heart rate extraction respectively (denoted as the labels of RR/HR hereafter). During the manual labeling process, the availability is determined based on signal patterns, such as how periodic and continuous the signal is in the time domain, and/or how condensed the spectrum is in the frequency domain around the frequencies of RR/HR. It takes over 80 hours to annotate the labels of RR and HR for all 50268 samples in data set $D1$. The manual labeling process is repeated twice (each takes over 40 hours) to mitigate occasional label errors due to subjective mistakes. After such repeated manual labeling, we confirm that the label errors are mostly eliminated by examining randomly selected samples: the averaged error ratio has been reduced from above 10% to less than 1% based on 5 batches of 200 randomly selected samples.

Interestingly, we note that the labels of RR and HR often differ from each other, which is a phenomenon not identified in existing literature. In data set $D1$, we observe only 71% of the labels of RR agree with the labels of HR. To quantify the level of agreement between the labels of RR and HR, we use Cohen's Kappa (κ) [17], which is a more robust measure than simple percent agreement calculation, as it takes into account the possibility of the agreement due to randomness:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}, \tag{6}$$

where $p_o = \frac{1}{N}\sum_k n_k^{RR,HR}$ is the observed agreement between the labels of RR and HR, $p_e = \frac{1}{N^2}\sum_k n_k^{RR} n_k^{HR}$ is the hypothetical probability of chance agreement [17, 31], $n_k^{RR,HR}$ denotes how many times the labels of both RR and HR are assigned to the same class $k \in \{0\ for\ unavailable, 1\ for\ available\}$, N is the total number of samples to inspect, n_k^{RR} and n_k^{HR} denotes how many labels of RR and HR are assigned to class k respectively.

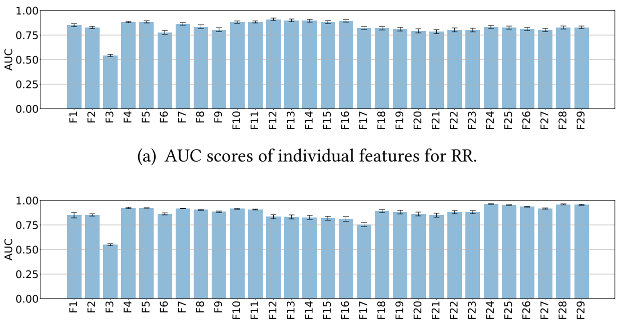
The higher the Cohen Kappa (κ), the more agreement between two labels, and the value of 1 indicates a complete agreement. The calculated κ of 0.48 indicates a moderate agreement [22] between the labels of RR and HR. This necessitates two separate SQDs for RR and HR. Since they together produce 2 bits to indicate the respective Boolean states of the availability of RR and HR, we name our method "*2-bit SQD*".

4.2.2 Feature Extraction. The physiological movements (i.e., respiration and heartbeat) being continuous and periodic is a critical pattern to distinguish available signals from unavailable ones. A list of feature extractors are devised to extract discriminative features (as listed in Table 1) from the stored data to describe critical patterns in both time and frequency domains.

Table 1: Feature $F1$-$F29$ are extracted with different combinations of feature extractors and inputs in the time and frequency domains.

Input	Feature Extractor							
	Rie	*PAR*	*mPAR*	*PNR*	*Var*	*Ent*	*Ske*	*Kur*
T	$F1$				$F2$		$F3$	$F4$
$S0$		$F5$	$F6$	$F7$	$F8$	$F9$	$F10$	$F11$
$S1$		$F12$		$F13$	$F14$	$F15$	$F16$	$F17$
$S2$		$F18$		$F19$	$F20$	$F21$	$F22$	$F23$
$S3$		$F24$		$F25$	$F26$	$F27$	$F28$	$F29$

- *Riemann sum* ("*Rie*"). The Riemann sum of the absolute values of time series is a critical indicator of the continuity [23], which is a common pattern of the available signal; while the corrupted signal likely has large spikes.
- *Peak-to-average Ratio* ("*PAR*"). The peak-to-average ratio of the power spectrum in the frequency domain indicates the sharpness of the spectrum. A periodic signal in the time domain translates to a spectrum with condensed energy in limited frequency bands, while a noisy signal spreads the energy over the entire spectrum.
- *Multi-peaks-to-average Ratio* ("*mPAR*"). Because the perceived signal reflected from the chest wall is modulated by a superposition of respiration and heartbeat, there exist multiple peaks in several different spectral bands. The ratio of multiple such peaks to the average can indicate the signal availability. Empirically, we use the sum of peaks whose magnitude is above the 75-percentile of the spectrum to calculate the ratio of multiple peaks to the average.
- *Peak-to-noise Ratio* ("*PNR*"). We find "*PAR*" and "*mPAR*" may not be representative when there exist small jitters in the power spectrum of the available signal due to intermodulation effects between respiration and heartbeat. To supplement such cases, we use the peak-to-noise ratio, calculated with the averaged floor of the power spectrum, which is empirically set to the lowest 25% of the power spectrum.
- *Variance* ("*Var*"). In the time domain, the corrupted signal with large spikes will have larger variance than the available signal with continuity; in the frequency domain, the corrupted signal will have a flat spectrum, thus relatively smaller variance than available signals with energy condensed around the frequencies of vital signs in the spectrum.



(a) AUC scores of individual features for RR.

(b) AUC scores of individual features for HR.

Figure 3: The difference in AUC scores between RR and HR of each individual feature indicates that the same feature is not equally effective in detection of RR and HR. Features $F12$-$F16$ are more effective in detection of RR, while features $F24$-29 are more effective in detection of HR.

- *Entropy* ("*Ent*"). Inherently, Shannon's Entropy indicates the level of randomness. Unavailable signals from disturbances have more randomness than available ones with periodic vital signs.
- *Skewness* ("*Ske*"). Skewness is a measure of the asymmetry of the distribution with the third standardized moment. The corrupted signal will have more asymmetric distribution in the time domain due to larger spikes, and more symmetric distribution in the frequency domain due to flatter spectrum.
- *Kurtosis* ("*Kur*"). Similarly, kurtosis describes the "tailedness" of a probability distribution from outliers using the fourth standardized moment. The corrupted signal in the time domain has large spikes, which add to outliers in the distribution, thus lower value of kurtosis. The corrupted signal in the frequency domain appears flatter as its energy spreads over the whole spectrum, which results in a higher kurtosis compared to the available signal with condensed energy around the frequencies of vital signs.

Each UWB data sample is a time series (denoted by "T") of the phase values in the echo from the chest wall in a 30-second window. The phase value is sampled at 10 Hz , and the window moves at 1-second increments. According to the Nyquist–Shannon sampling theorem [20], the Fourier transform (given 10 Hz sampling rate) produces a spectrum in the frequency range of 0–5 Hz, which translates to a physiological rate range of 0–300 beats/breaths per minute (bpm). The original spectrum within the full range of 0–300 bpm is denoted by "$S0$". Besides, the spectrum within the range of RR 0–30 bpm is denoted by "$S1$", the spectrum within the range of the fundamental HR 50–150 bpm denoted by "$S2$", and the spectrum within the range of the second harmonic of HR 100–300 bpm denoted by "$S3$". Since there exists disagreement between the availability of RR and HR, we apply feature extractors to extract features in respective spectrum ($S1$, $S2$ and $S3$) specific to RR/HR, in addition to the original time series (T) and the full spectrum ($S0$). Table 1 shows "combos" of feature extractors and inputs to generate features.

4.2.3 Effectiveness and Redundancy of Features. To understand the effectiveness of the extracted features for SQD, we use each feature individually to see how well it can distinguish between available and corrupted signals, using data set $D1$ and the AUC metric [8]. The AUC score usually ranges in value from 0.5 to 1. The higher the AUC score, the more discriminative the corresponding feature is to distinguish between available and corrupted signals; while an AUC score of 0.5 indicates that the feature has no discrimination capacity but random prediction. We apply 5-fold cross validation to mitigate the variance due to data distribution [14]. Figure 3(a) and Figure 3(b) show AUC scores of individual features for detection of RR and HR. Notably, every individual feature performs differently for the detection of RR and HR, and two subsets ($F12 - 16$, $F24 - 29$) seem more effective for RR, HR respectively. Therefore, we use different subsets of features to devise SQDs for RR and HR.

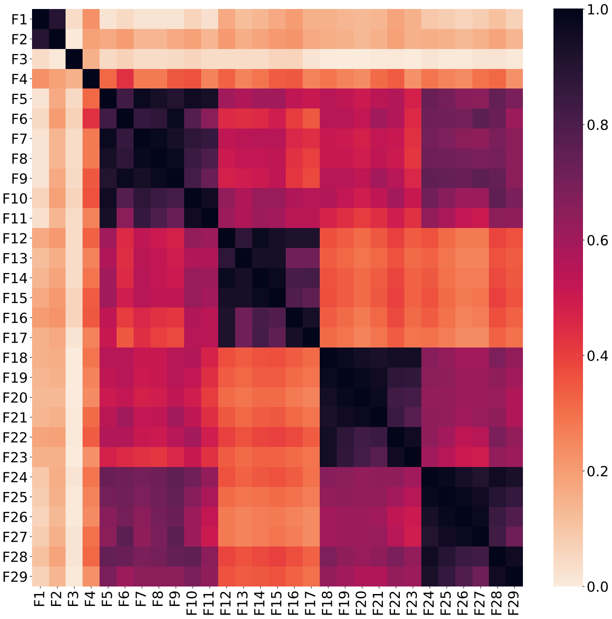
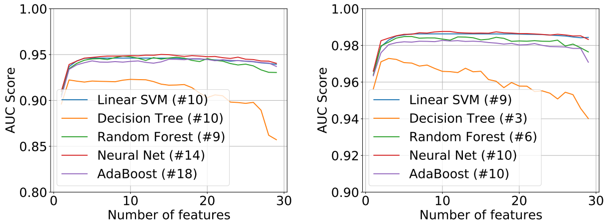


Figure 4: The correlation between pairs of features F_i and F_j is indicated by the cell located in the coordinate (F_i, F_j). The darker a cell is, the more correlated the corresponding pair of features are.

We also evaluate the redundancy between different feature variables using Pearson correlation [4]. The Pearson correlation coefficient between feature variables F_i and F_j is calculated as:

$$\rho_{i,j} = \frac{\mathrm{Cov}(F_i, F_j)}{\sigma_i \sigma_j},$$

where Cov is the covariance, σ_i and σ_j are standard deviations of F_i and F_j respectively. The coefficient value is between -1 and 1. The value of 1 indicates positive correlation between the pair of features; the value of 0 indicates totally no correlation; and the value of -1 indicates negative correlation. We use the absolute value of the coefficient because both positive and negative correlations indicate a certain level of redundancy between the pair of features. The resulting correlation matrix is shown in Figure 4. Interestingly, higher correlation is observed among features extracted from the same input (e.g., features $F5$–$F11$ extracted from $S0$; $F12$–$F17$ from $S1$; $F18$–$F23$ from $S2$; and $F24$–$F29$ from $S3$), indicating such features contain more redundant information. Including more redundant features will not contribute to higher classification performance,



(a) Forward selection for RR.

(b) Forward selection for HR.

Figure 5: The curves show the changes in AUC scores of classifiers along the step-wise forward selection procedure with increasing number of features for detection of RR in (a) and HR in (b). The AUC score increases marginally after 2 or 3 features, until it reaches the peak with the number of features as indicated in the label of each classifier.

but may introduce more noise and reduce the performance. We need to select optimal subsets from these features for 2-bit SQD.

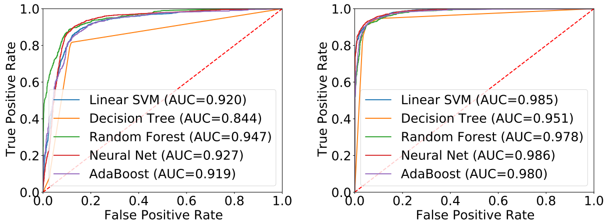
4.2.4 Feature Selection for Signal Quality Detection. We apply the step-wise greedy forward selection strategy [9, 15] to evaluate the contributions of features and select an optimal subset of features that achieves the best detection performance. The forward selection method starts by choosing the first feature that produces the best AUC score [8] among all feature candidates. Then, among the remaining candidates, the next best feature that contributes the most increase of AUC score when combined with selected features, is selected. We repeat the above step and progressively add the more features to the subset of selected features, until we exhaust all remaining feature candidates.

We empirically choose to investigate 5 popular machine learning models as the classifiers of our 2-bit SQD. They are Support Vector Machine with linear kernel [7] (denoted as "Linear SVM"), Decision Tree [29], Random Forest [25], Neural Net (configured as a 3-layer perceptron classifier with the activation function of ReLu) [10], and AdaBoost [27].

Figure 5(a) and Figure 5(b) show the changes in AUC scores as more features are selected. For all classifiers, we consistently observe that the AUC scores rise with the selection of more features at first. Then it increases only marginally after 2 or 3 features, until it reaches the peak with the number of features as indicated in the label of each classifier (from 3 to 18).

There are two reasons behind the drop in AUC score with further increasing numbers of features: 1) redundant features add little new information but mostly noise detrimental to performance; 2) with more features, a machine learning model needs to tune more parameters during training, and it becomes more likely to be under-trained given the limited size of training data.

We will select the respective optimal feature subset for each classifier when the best AUC score is achieved. It is noted for most classifiers, $F12$ and $F28$ are two most effective features for classification of RR, while $F24$ and $F4$ are two most effective ones for HR. It implies that, while the features extracted from the respective spectrum bands of RR/HR are most effective, features with less correlation contribute more to supplementary information thus increments in performance.



(a) ROC comparison for RR.

(b) ROC comparison for HR.

Figure 6: ROC curves of different classifiers with respective optimal feature subsets are close to each other for either RR in (a) or HR in (b). It shows that with carefully selected features, the impact of difference in the classification algorithms is limited.

While it is possible to choose a smaller feature subset (e.g., the first 3-5 features that already approaches optimal AUC) with a marginal loss in classification performance, we use the identified optimal feature subsets (most are with 3-10 features) for each classifier and train them as SQDs. We will evaluate these SQDs for their end to end performance next.

5 EVALUATION

In this section, we first conduct experiments to evaluate the classification performance of the SQDs in §5.1 on the testing data collected from unseen subjects ($D2$), then we evaluate the end-to-end performance of vital sign monitoring in §5.2 on the data collected over 3 months ($D3$). In all experiments, we use the classification models trained on data $D1$, with the subset of features selected in §4.2.4.

5.1 Classification Performance

We evaluate the classification performance on the data set $D2$, which contains subjects unseen in the training data $D1$. Figure 6(a) and Figure 6(b) show the ROC curves of respective trained classifiers to detect RR and HR in data $D2$. We observe that the ROC curves of different classifiers are near each other with very close AUC scores (mostly over 0.91), despite using different feature subsets. This shows that with careful feature selection, the classifiers can all perform similarly well for unseen subjects.

We also evaluate the gain achieved by consensus filtering [16] (denoted by "*Consensus*"), which takes advantage of the diversity in classifiers and classifies a sample as "available" only when all classifiers says so. We compare the performance of 2-bit SQD with two representative methods from the literature – the flat spectrum detector (FSD) [3] and Spectrum-averaged Harmonic Path detector (SHAPA) [24]. Notably, all existing methods are 1-bit solutions, which assume RR and HR are either both "available" or "unavailable" concurrently. Since existing methods are not open-sourced, we try our best to implement them for comparison.

FSD is a threshold-based method, which uses the peak-to-average ratio to quantify the sharpness of the spectrum to indicate the periodicity of the signal. Varying the threshold value can cause different classification performances. We use the threshold when the best F1-score is obtained with the data set $D1$. SHAPA is a heuristic-based method, which determines the signal availability according to the existence of "harmonic paths" due to the inherit non-linearity characteristic. A harmonic path is detected if there exist three or more near equally spaced spectral peaks, whose frequencies are approximately an integer multiple of the inter-peak distance in terms of frequency [24]. The peaks with magnitudes above the 75-percentile of the spectrum are selected according to [24]. While it was proposed for both the detection and estimation of vital signs, we only compare its detection performance with our proposed SQD.

Table 2: The classification performance of trained classifiers are much improved compared to existing methods FSD and SHAPA.

Detector	RR		HR	
	Prec.	Rec.	Prec.	Rec.
Linear SVM	0.87	0.86	**0.96**	0.91
Decision Tree	0.87	0.82	0.93	0.91
Random Forest	**0.91**	**0.88**	0.95	0.91
Neural Net	0.89	**0.88**	**0.96**	**0.92**
AdaBoost	0.88	0.84	0.95	**0.92**
Consensus	**0.93**	0.80	**0.96**	0.91
FSD	0.83	0.79	0.85	0.80
SHAPA	0.51	0.22	0.63	0.37

Table 2 shows the classification performance tested on data $D2$ with metrics of *Precision* (i.e., $\frac{TP}{TP+FP}$, denoted by "Prec.") and *Recall* (i.e., $\frac{TP}{TP+FN}$, denoted by "Rec"). Precision indicates the fraction of truly "available" samples in those detected as "available." A high precision is necessary to ensure the estimation uses mostly truly available signals, thus low erroneous measurements. Recall is the fraction of detected available samples among all truly available ones. A high recall ensures most truly available signals are included for estimation, thus high coverage over time.

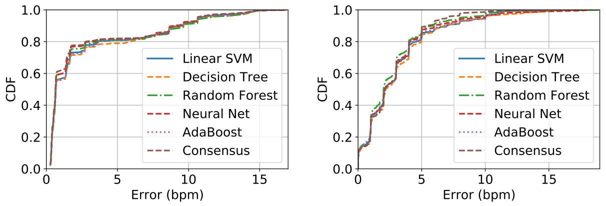
The performance of the existing methods may differ from the reported results in the original paper, as the data source is different. The performance of FSD is relatively reasonable (0.79-0.85) while not as good as the SQDs with optimal feature subsets (mostly around 0.90 or higher). However, SHAPA is much worse (0.22-0.63) than others. It is because the harmonic path exists only for a small fraction of available signals, leading to a very low recall (i.e., a small fraction of the true positive (TP) samples). This stringent criteria of availability also reduces the amount of false positive (FP). But if its level is comparable to TP, thus the precision ($\frac{TP}{TP+FP}$) is still low.

The difference in classification performance among our trained SQDs is marginal as shown in both Figure 6 and Table 2. Using consensus filtering over predictions of all classifiers further improves the precision marginally at the cost of decreased recall. The classification results show that the proposed 2-bit SQD is robust to unseen data, and achieve much improved performance than existing methods.

5.2 End-to-end Performance

To evaluate the impact of the SQDs on the end-to-end vital sign estimation performance, we use the longitudinal data $D3$ which contains more variations over 2 months' period.

The RR and HR estimations are produced using the same vital signs estimation method (which has been extensively evaluated in [33], and its detailed design is beyond the scope of this paper), while only changing SQDs. Only the readings of RR/HR classified as available by the SQD will feed the estimation method, and the resulting vital signs data are compared against the ground truth with the metric of errors in "beats per minute" (bpm).

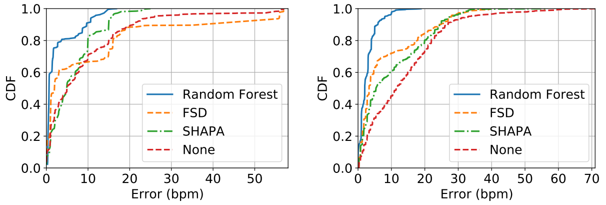


(a) E2E estimation performance of RR. (b) E2E estimation performance of HR.

Figure 7: CDF curves of E2E performance with different SQDs are close to each other.

Table 3: E2E vital signs estimation errors (bpm) at 80-percentile with different SQDs. The difference in the E2E performance with different SQDs is marginal.

Detector	RR	HR
Linear SVM	3.72	4.61
Decision Tree	5.72	4.97
Random Forest	3.51	**4.01**
Neural Net	3.41	4.13
AdaBoost	3.72	4.29
Consensus	**2.83**	**4.01**



(a) E2E estimation performance of RR. (b) E2E estimation performance of HR.

Figure 8: E2E performance comparison of our random forest based 2-bit SQD with existing work FSD, SHAPA, and without any SQD.

We first compare the end-to-end performance using different classifiers with respective optimal feature sets, as well as the consensus filtering. As shown in Figure 7, all the classifiers achieve similar performance, due to close scores of precision and recall in Table 2. The errors in Table 3 show that 80-percentile errors vary between 2-5 bpm, with the best for RR and HR around 2.83 and 4.0 bpm, respectively.

While the consensus filtering is only marginally better than Random Forest in the E2E performance of RR, it consumes predictions from all 5 classifiers, which is (about 5 times) more complex compared to a single classifier. To balance between complexity and performance, we decide to only keep the trained Random Forest for 2-bit SQD. The E2E RR/HR monitoring performance using our proposed 2-bit SQD is then compared with two existing approaches (FSD [3] and SHAPA [24]) as well as results without any SQD.

Figure 8 shows the results of different CDF curves, which clearly shows our 2-bit SQD achieves the best results, with much reduced overall errors for both RR and HR (i.e., the curve is above others, with smaller errors on X axis at any given percentile on Y axis). Using Random Forest with the derived feature subsets as the 2-bit SQD reduces the 80-percentile RR/HR error to 3.5/4.0 bpm, from 15/22 bpm of the baseline without SQD. Compared to the 80-percentile error of 16/18 bpm with FSD and 10/20 bpm with SHAPA, our proposed 2-bit SQD achieves about 3~4-fold reduction in E2E RR/HR errors compared to the better one between FSD and SHAPA. The E2E results show that, our proposed 2-bit SQD can reliably detect available signals thus controlling the measurement errors of RR/HR, which makes it more practically useful for longitudinal vital sign monitoring in home.

6 DISCUSSION

Practical Concerns of Wireless Sensing. An early study [6] found that exposure to RF has no detrimental effects on health. Our system targets longitudinal in-home vital signs data collection, and we introduce 2-bit SQD to exclude disturbed signals to avoid erroneous vital signs measurements while keeping the valid ones. The fact that people spend the majority of time in sedentary behaviors, as indicated in [21], allows our system to collect valid vital signs data when they stay stationary over a reasonably long duration for tracking the daily trends. We will investigate the medical usefulness in the future work with analysis over long-term records of vital signs.

Strength and Generalizability of 2-bit SQD. As 2-bit SQD allows decoupled detection of RR and HR, it is inherently more flexible and robust than other single-binary detection methods when there exists disagreement between the availability of RR and HR. We use a commodity UWB sensor as the RF front end in our wireless sensing system. However, the issues discussed in this paper (e.g., disturbances, intermodulation) are common to other wireless techniques (e.g., FMCW, mmWave). In future, we will evaluate the effectiveness of the proposed 2-bit SQD to other wireless sensing platforms.

Trade-offs between Complexity and Performance. In this paper, we use the subset of features with the best AUC score. However, it is noted that, before it reaches the best AUC, the increase in AUC score is actually marginal after the selection of the first 2 or 3 features. This may further reducing computing complexity at the cost of hopeful sligtly lower classfication accuracy. We will leave it to future work to find a proper balance between the two based on larger amounts of data collection.

Further Improvement in Accuracy. While the idea of 2-bit SQD is verified, we use mostly statistical features in time and frequency domains. We will explore more categories of features (e.g., local patterns using wavelet analysis, representations learned from neural networks) classification algorithms to further improve the performance of SQD. Although we control of error of RR/HR measurement at 80-percentile to a reasonable level (3.5/4.0 bpm), the error beyond 80-percentile is still large. Since it is impractical to eliminate false positive samples completely, we will leverage prior knowledge about the past history and trend of vital signs (e.g., continuity using LSTM [12]) to supplement SQD for more reliable measurement of vital signs.

More Subjects for Data Diversity. The diversity of our data is limited due to the COVID-19 pandemic. We were able to only recruit a limited number of participants. As the situation eases with warmer weather and vaccination, we will further recruit more participants to increase the diversity of data, possibly including residents in nearby communities for more diversities in age, health status, and body build.

7 CONCLUSION

We propose a 2-bit signal quality detector (SQD) for robust wireless non-touch vital sign monitoring. We identify and quantify the disagreement of the availability between RR and HR signals, breaking an implicit assumption made in all prior work. We propose to treat RR and HR separately, examine a rich set of time and frequent domain features, and identify the optimal feature subset for a number of common classification algorithms. Experiments show that our proposed 2-bit SQD improves signal quality detection, and achieves about 3~4 fold reduction in E2E RR/HR errors at 80-percentile compared to existing work.

REFERENCES

[1] [n. d.]. Masimo. https://www.masimo.com

[2] [n. d.]. X4M03. https://www.xethru.com

[3] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. 2015. Smart homes that monitor breathing and heart rate. In *ACM CHI'15*.

[4] Jacob Benesty, Jingdong Chen, and Yiteng Huang. 2008. On the importance of the Pearson correlation coefficient in noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 4 (2008), 757–765.

[5] Lynn Bickley and Peter G Szilagyi. 2012. *Bates' guide to physical examination and history-taking*. Lippincott Williams & Wilkins.

[6] David R Black and Louis N Heynick. 2003. Radiofrequency (RF) effects on blood cells, cardiac, endocrine, and immunological functions. *Bioelectromagnetics* 24, S6 (2003), S187–S195.

[7] Nello Cristianini, John Shawe-Taylor, et al. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

[8] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[9] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.

[10] Mohamad H. Hassoun et al. 1995. *Fundamentals of artificial neural networks*. MIT press.

[11] Peter Hillyard, Anh Luong, Alemayehu Solomon Abrar, Neal Patwari, Krishna Sundar, Robert Farney, Jason Burch, Christina Porucznik, and Sarah Pollard. 2018. Experience: Cross-Technology Radio Respiratory Monitoring Performance Study. In *ACM Mobicom'18*.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[13] Omer T Inan, Mozziyar Etemadi, Richard M Wiard, Laurent Giovangrandi, and GTA Kovacs. 2009. Robust ballistocardiogram acquisition for home monitoring. *Physiological measurement* 30, 2 (2009), 169.

[14] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Montreal, Canada, 1137–1145.

[15] Ron Kohavi and George H John. 1997. Wrappers for feature subset selection. *Artificial intelligence* 97, 1-2 (1997), 273–324.

[16] Effrosyni Kokiopoulou and Pascal Frossard. 2010. Distributed classification of multiple observation sets by consensus. *IEEE Trans. Signal Process* (2010).

[17] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[18] Changzhi Li, Yanming Xiao, and Jenshan Lin. 2006. Experiment and spectral analysis of a low-power Ka-band heartbeat detector measuring from four sides of a human body. *IEEE T-MTT* (2006).

[19] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. 2015. Tracking vital signs during sleep leveraging off-the-shelf wifi. In *ACM MobiHoc*.

[20] Robert J II Marks. 2012. *Introduction to Shannon sampling and interpolation theory*. Springer Science & Business Media.

[21] Charles E Matthews, Kong Y Chen, Patty S Freedson, Maciej S Buchowski, Bettina M Beech, Russell R Pate, and Richard P Troiano. 2008. Amount of time spent in sedentary behaviors in the United States, 2003–2004. *American journal of epidemiology* 167, 7 (2008), 875–881.

[22] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[23] Robert R Muise and Charles K Chui. 1995. Wavelet approach to detect discontinuities of intensity functions for minefield classification. In *Detection Technologies for Mines and Minelike Targets*, Vol. 2496. SPIE, 531–542.

[24] Van Nguyen, Abdul Q Javaid, and Mary Ann Weitnauer. 2014. Spectrum-averaged Harmonic Path (SHAPA) algorithm for non-contact vital sign monitoring with ultra-wideband (UWB) radar. In *IEEE EMBC'14*.

[25] Mahesh Pal. 2005. Random forest classifier for remote sensing classification. *International journal of remote sensing* 26, 1 (2005), 217–222.

[26] Jiapu Pan and Willis J Tompkins. 1985. A real-time QRS detection algorithm. *IEEE transactions on biomedical engineering* 3 (1985), 230–236.

[27] Raúl Rojas et al. 2009. AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting. *Freie University, Berlin, Tech. Rep* (2009).

[28] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*. Ieee, 1297–1304.

[29] Yan-Yan Song and LU Ying. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry* 27, 2 (2015), 130.

[30] J Abdul Sukor, SJ Redmond, and NH Lovell. 2011. Signal quality measures for pulse oximetry through waveform morphology analysis. *Physiol. Meas.* (2011).

[31] Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med* 37, 5 (2005), 360–363.

[32] Xuyu Wang, Chao Yang, and Shiwen Mao. 2017. PhaseBeat: Exploiting CSI phase data for vital sign monitoring with commodity WiFi devices. In *IEEE ICDCS*.

[33] Zongxing Xie, Bing Zhou, Xi Cheng, Elinor Schoenfeld, and Fan Ye. 2021. VitalHub: Robust, Non-Touch Multi-User Vital Signs Monitoring using Depth Camera-Aided UWB. In IEEE ICHI'21.

[34] Moustafa Youssef, Matthew Mah, and Ashok Agrawala. 2007. Challenges: device-free passive localization for wireless environments. In *ACM Mobicom'07*.

[35] Li Zhang, Chuanwei Ding, Xudong Zhou, Hong Hong, Changzhi Li, and Xiaohua Zhu. 2020. Body movement cancellation using adaptive filtering technology for radar-based vital sign monitoring. In *IEEE RadarConf20*.